Procedia Computer Science

Volume 51, 2015, Pages 755–764

ICCS 2015 International Conference On Computational Science

# Using visual analytics to support the integration of expert knowledge in the design of medical models and simulations

Philippe J. Giabbanelli, PhD[1] and Piper J. Jackson, PhD[2]

[1] School of Clinical Medicine, University of Cambridge, United Kingdom
pg438@cam.ac.uk

[2] Interdisciplinary Research in the Mathematical and Computational Sciences (IRMACS) Centre, Simon Fraser University, Canada
pjj@sfu.ca

**Abstract**

Visual analytics (VA) provides an interactive way to explore vast amounts of data and find interesting patterns. This has already benefited the development of computational models, as the patterns found using VA can then become essential elements of the model. Similarly, recent advances in the use of VA for the data cleaning stage are relevant to computational modelling given the importance of having reliable data to populate and check models. In this paper, we demonstrate via case studies of medical models that VA can be very valuable at the conceptual stage, to both examine the fit of a conceptual model with the underlying data and assess possible gaps in the model. The case studies were realized using different modelling tools (e.g., system dynamics or network modelling), which emphasizes that the relevance of VA to medical modelling cuts across techniques. Finally, we discuss how the interdisciplinary nature of modelling for medical applications requires an increased support for collaboration, and we suggest several areas of research to improve the intake and experience of VA for collaborative modelling in medicine.

*Keywords:* Data visualization; Knowledge management; Medical models

## 1 Introduction

Modelling is the process of identifying the common properties of closely related phenomena or entities and then using these to produce a useful characterization. In the case of computational modelling, this utility is expressed in the construction of models that can be run as a software program on a computer. Rossiter *et al.* highlight an important distinction between the intended uses of a model: theoretical, exploring or expressing a concept; or descriptive, simulating the behaviour of a specific real-world scenario [16]. In this paper, we focus on the development of descriptive models to support medical decision-making. Consequently, the real-world scenarios studied here lead to models where medical data, health outcomes and health behaviours play an

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015


doi:10.1016/j.procs.2015.05.195

important role. In this setting, computational modelling involves several tasks: (i) data cleaning and data analysis, (ii) choice of modelling technique(s) and construction of the model(s), (iii) experimentation via computer simulations, and (iv) reporting to stakeholders.

Visual analytics has been increasingly used for data analysis and reporting to stakeholders. A growing demand to integrate visual analytics with these tasks has been met with the development of numerous off-the-shelf software packages. For example, we started by analyzing a dataset about drinking behaviours in a Dutch sample via classical data mining methods [2]. In a second phase, we integrated the software `Tableau` into our analysis, and we used interactive visualizations to assess what factors had a marked difference between binge drinkers and non-binge drinkers. The analysis (Figure 1) revealed that social factors consistently displayed clear behavioural differences between binge drinkers and non-binge drinkers. This use of visual analytics led us to focus on modelling social factors in drinking, which was ultimely successful in explaining the behaviour of most of the sample [4]. The dissemination and reporting of research results to stakeholders is also facilitated by many applications, such as `Spotfire`'s DecisionSite Posters, in which interactive analysis reports can be viewed in a Web browser. Similarly, numerous websites have supported ways to share and get feedback on visualizations, such as `Many-Eyes.com` (e.g., comments can be left on the visualizations) or `sense.us` where specific visualization states can be commented upon [8].

In contrast with data analysis and the dissemination of results, the integration of visual analytics with the other tasks of computational modelling remains limited. The landmark article by Kandel et al. [10] provided numerous examples and research directions for integration with data cleaning and wrangling. An array of tools has been developed to facilitate the use of visual analytics for simulation results. For example, several tools have been presented to simulate and explore large datasets in the context of the `IMAGE` project [13]. This paper's main contribution is on the central and relatively unexplored theme of integrating visual analytics with the design of a computational model.

### 1.1 Contribution of the paper

The principal contribution of the present work is to provide practical ways to integrate visual analytics with the design of a computational model for medical decision-making. Specifically, we draw on three case studies to show how the conceptual stage of the design process (rather than the mathematical or implementation stages) can be improved by using visual analytics. This is demonstrated for different stages of conceptual modelling, and the benefits of such an integration at each stage are highlighted.

### 1.2 Organization of the paper

In section 2, we situate the task of conceptual modelling within the overall modelling pipeline. In section 3, we present the case studies regarding the integration of visual analytics in the design of conceptual models. Finally, we discuss how the interdisciplinary nature of biomedical research requires an increased support for collaboration, and we provide specific areas for future research in visual analytics to address this need.

## 2 Background

Generally the process of simulation modelling can be thought of as taking part in 4 general phases:

(1) Data acquisition and analysis, where the relevant data for the subject is explored
(2) Modelling, where the model is specified and constructed
(3) Experimentation, where the model is used to produce results
(4) Reporting, where discoveries and knowledge are shared


Figure 1: A Dutch sample (n=2837) completed a questionnaire about drinking based on 4 categories (conformity, coping, enhancement, social), with 5 questions per category. The reference line (1.0) shows answers that were endorsed at the same proportion by binge drinkers and non-binge drinkers. Answers above this line are more frequent for binge drinkers. Questions on social aspect triggered the more marked difference between binge drinkers and non binge-drinkers, with a consistent gradient among answers from High (H) to Very Low (VL).

In descriptive models, *phase 1* is the spark: the identification of key patterns in the relevant data may provide the intuition or inspiration for modelling. Visual analytics is increasingly used at that phase, with a plethora of off-the-shelf software such as `in-Spire` or `STARLIGHT` (qualitative data), `GeoTime` (spatio-temporal data), `Gephi` (network data), or `Tableau` (multi-dimensional relational databases). There are also specific tools answering to the different types of data found in the medical world (*e.g.*, genomic, biomarkers). For example, the newly launched `Sequence Bundles` has specifically been designed for multiple sequence alignments [11].

The modelling process in *phase 2* can be further divided. Since modelling itself is a process containing common tasks and concerns, there are several models of the modelling process. They typically share a depiction of modelling as a highly iterative, complex process, where adjustments and decisions are made based on feedback between the different stages of development. However, they differ in the aspects that they emphasize. For example, Sargent describes all 4 phases outline above but focuses on the verification and validation of the model [18]. This leads to an emphasis on the interactions between the simulated system and its real world counterpart, such as by comparing data (gathered in the real world vs. simulation experiment results), or by generating hypotheses (based on observation of the real world or the model, and tested on its counterpart). In this section, we follow the division of the second phase proposed by Brantingham *et al.* [1] and refined in [9]. This division has three stages: conceptual (a human description of the model), mathematical (a formal description of the model), and finally the implementation into a computer program. All three stages can benefit from visual analytics: for

example, `NDepend` already provides support to integrate visual analytics at the computational modelling stage. This paper focuses on the conceptual stage (section 3).

*Phase 3* uses the computer program to generate data. While the data used to develop a model can be analyzed via the tools aforementioned, much fewer tools are developed to analyze the data generated by a model. The `IMAGE` project [13] provides an example of such a tool, although its focus is on visual analytics for simulations in defense rather than in the medical realm. Finally, *phase 4* converts the ideas and experiences of the research team into knowledge that can be shared with the stakeholders and/or other researchers.

This paper focuses on the use of visual analytics at the conceptual stage of the modelling process. The conceptual stage can itself be further divided based on how advanced it is. The next section details these divisions, and provides the benefits of visual analytics for each of them in the case of medical applications.

# 3 Visual analytics at the conceptual stage

In this paper, we divided the conceptual stage of the modelling process based on when visual analytics was used. This led to three situations, each presented through a case study. In the first situation, visual analytics was used early on: it served to provide the foundations of the conceptual model. This is illustrated through the design of a new simulation model for obesity, which aimed at integrating both the physiological and psychological factors related to obesity. The second situation is at the other extreme, where visual analytics was used late in the process: the conceptual model was already developed and considered final. Visual analytics thus served to assess the fit of the model with the underlying data. This is exemplified through a large international program for health behaviour change, where the conceptual model synthesized the views of different stakeholders. The third situation sits between the two extremes: it uses visual analytics and conceptual modelling in an iterative way. The model is refined through exploration of the data, which in turns provides better guidance to the exploration, which further improves the model. Due to space constraints, the main principles of this iterative refinement will be exposed and we will refer the reader to additional material.

## 3.1 Using visual analytics at an early stage

In 2014, we were mandated to create a simulation model of obesity, which would include psychological as well as physiological factors. We used a discussion paper published by the Provincial Health Services Authority (PHSA) of British Columbia as a point of departure [15]. Analyzing the relationships in this paper was deemed necessary for the development of the simulation model for two reasons. First, it could provide an insight into the dynamics of obesity [5, 6]. Second, understanding it would strengthen the foundations to gradually develop more comprehensive models of obesity. However, the structure of conceptual models of obesity has been typically analyzed by computing summary statistics (e.g., using Pajek [3]) rather than through the use of visual analytics. In this case study, we used visual analytics to examine the conceptual model created in the PHSA document and thereby create the requirements for the development of the next conceptual model.

We manually extracted all relationships mentioned in the PHSA report in order to obtain network data. Each of the factors involved in a relationship was assigned a category (Figure 2(a)), using categories similar to those in the Foresight Obesity Map [19]. We first used the visual analytics tool `Gephi` to analyze the map with respect to these categories. The hypothesis was that factors within the same category should be interacting.

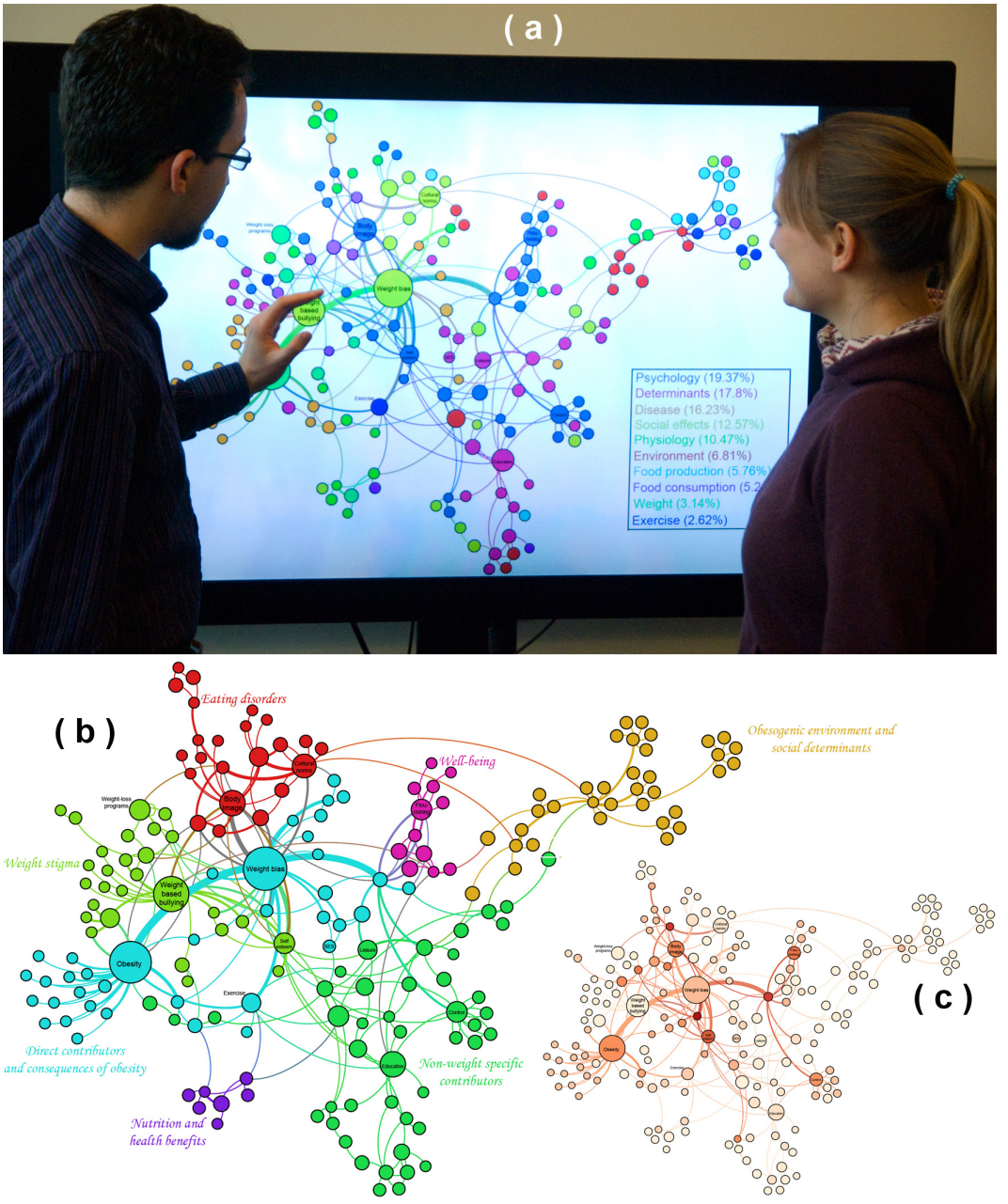


Figure 2: The map consisting of the factors and relationships mentioned in the PHSA's discussion paper [15] was analyzed by (a) coding factors in the same way as in the Foresight Map [19], (b) creating modules that better represented the uniqueness of this report, and (c) the structural centrality of its different factors. This was done in an interactive manner (a).

For example, one could expect factors from food production (in cyan on Figure 2(a)) to be clustered in a similar way to factors from food consumption (in purple Figure 2(a)). Visual analytics quickly led to the dismissal of this hypothesis, indicating that new categories had to be developed to understand the previous conceptual model. We then used visual analytics to explore possible ways to group factors (*i.e.*, to create modules in the map) and obtained a set of 7 modules whose factors were considered related by the research team. We analyzed these modules as well as the role that they played (*e.g.*, to assess what seems most central in the model; Figure 2(c)) and we found several gaps. These gaps were turned into a roadmap for the development of the new model. For example, Figure 2(b) shows that there are direct consequences of obesity (*e.g.*, co-morbidities) but these do not feed back to the model. In reality, co-morbidities of obesity have several impact that should be captured: for example, issues in the cardiovascular system may impair one's ability to exercise and well-being. The roadmap created by analyzing the PHSA report through visual analytics was then put to action by teaming up with experts in fields such as adipose tissue and the cardiovascular system.

Generalizing the use of visual analytics to the early stage of the development of a conceptual model has the following advantages. First, it helps setting up the structure of the model by facilitating the identification of similar entities or missing connections between entities. Second, by grounding the development of the model on a sound structure, it contributes to eliminating wasteful development work that may otherwise occur later on. Finally, it contributes to keeping stakeholders engaged in the modelling process by providing an interactive outline of the domain that is being modelled such that new questions may be generated.

## 3.2 Using visual analytics at the late stage

The Mind, Exercise, Nutrition, Do it (MEND) program is an intervention for children with obesity and their families. The program consists of sessions on nutrition and behaviour change, as well as group physical activity [17]. While MEND originates from the UK, it has also been implemented in Australia, Canada, Denmark, New Zealand and the US. The program is designed so that it can be implemented in the community by people who are not health professionals. Due to the scaling-up of the program, both in its geographic reach as well as the number of participants, it is of particular interest to understand the complexity of the organizational structure. A conceptual model of that structure was developed by synthesizing the viewpoints of several stakeholders into a causal loop diagram (Figure 3(a)). The version of the diagram illustrated here is not final. Indeed, additional work is required such that the dynamics suggested by the diagram match the behaviour of the system expected by the stakeholders. In order to capture the vision of the stakeholders, interviews were performed and transcribed by the Chronic Disease Systems Modeling laboratory at Simon Fraser University, which provided ethics approval. In this case study, we used visual analytics to assess the fit between the diagram and the visions expressed by the stakeholders.

A total of 18 stakeholder interviews were produced by the Chronic Disease Systems Modeling laboratory. Interviews were divided at the level of answers; that is, the unit of analysis in this case study was the set of all sentences expressed until the interviewer's next question. For each relationship suggested in the causal diagram, we counted the number of answers that included both endpoints of the relationship. For example, the support for the relationship from 'program resource' to 'program implementation' was assessed by counting the number of answers that included both keywords. We accounted for common variations, which in this example include 'to implement', 'implementing', 'implementations'. The number of answers for each relationship is summarized in Figure 3(b), where variations of the keywords are suggested with *. Contrasting Figure 3(b) with the causal map in Figure 3(a) provides a valuable insight. While some relationships had a large amount of support, others had surprisingly little support,

such as the concept of 'program resources' which is only weakly linked to either 'program development' or 'program implementation'.

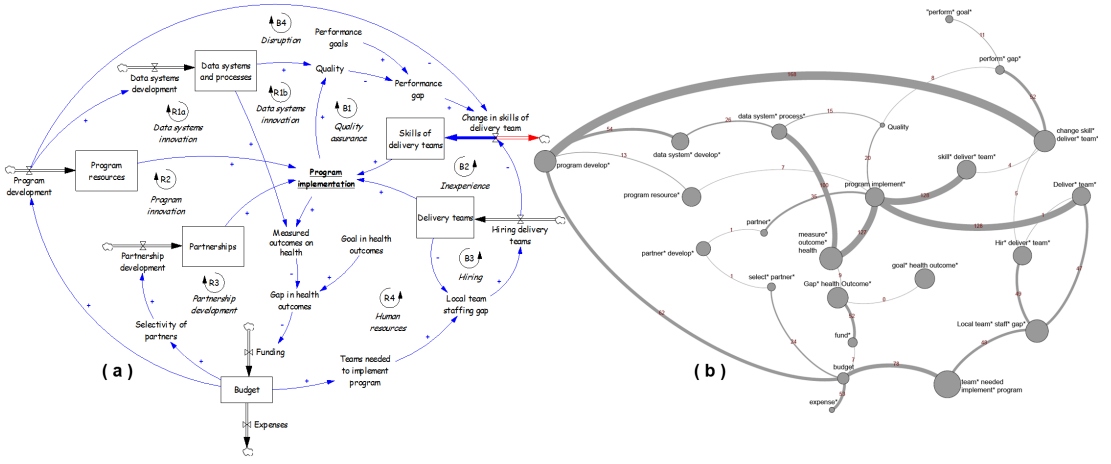


Figure 3: Causal map proposing an early theoretical model of the MEND program from the views expressed during meetings (a). *This map is the sole intellectual property of Dr. Peter Hovmand. It is reproduced with his permission.* In our visualization, the thickness and label of each edge is the number of times that the relationship was in an interviewee's answer (b).

This kind of examination provides three immediate benefits for the process of model building. First, it helps to delineate concepts. If 'program resources' is weakly linked, it could be that this concept doesn't appropriately cover the experience of the stakeholders. Consequently, the notion of 'resources' may have to be refined, possibly by merging it with either the 'budget' or the 'delivery team', both of which had better support. Second, it can identify gaps in data collection: providing Figure 3(b) to the stakeholder can lead to the realization that some relationships were actually important, but that none of the interviews in the sample covered it. Finally, generalizing this procedure to all pairs of interactions (rather than the ones suggested in the causal map) can uncover strong support for missing relationships, which could then be accounted for in the map, thereby making it more comprehensive.

## 3.3 Using visual analytics at an intermediate stage

The creation of a model for medical decision-making can be a process spanning several years, as numerous experts have to be consulted and heterogeneous data sources may have to be linked. During this time, more and/or better data may become available. For example, a model may have initially been designed on the release 5 of a dataset having blood samples for 5234 patients. Three months later, release 6 of the dataset may become available with 7320 patients, where some previously measured biomarkers have now been measured using more robust tools. Similarly, a model may be developed for a medical phenomenon that changes as it is being modelled, such as the Ebola epidemic in Africa.

On the one hand, visual analytics allows the interactive visualization of the new data, thereby supporting the modellers in finding the key patterns for the object of study. However, the complexity and large size of this data can make it difficult to efficiently navigating it, pointing to the need for a guiding framework. On the other hand, conceptual models provide that guiding framework but can gradually lose their relevance as either the dynamics change or

a better understand of these dynamics is obtained. For these reasons, we previously suggested that conceptual models and visual analytics should be coupled. This coupling uses conceptual models to enhance the traditional approach of visual analytics by providing a structure that supports the analyst in navigating a large search space in a focused way. This navigation will in turn lead to the discovery of new patterns, or discrepencies between the model and the framework, leading to an update in the model; we refer the reader to [14] for technical details.

There are several ways to operationalize this coupling, depending on the extent to which visual analytics is guided by a model. At the simplest level, this coupling can readily be integrated into current practices in `STARLIGHT`, where conceptual model can be used as a hierarchy. At the top-level, the model provides the key themes relevant to the domain. At the intermediate-level, the model specifies how to detect these themes, for example through a list of keywords possibly combined with a thesaurus to search for variants. Newly acquired data constitutes the bottom-level. To make sense of that data, modellers can select the top-level themes, which will permeate through the intermediate-level and eventually select relevant documents.

# 4 Discussion

In this paper, we presented how visual analytics can benefit modelling for medical decision-making by drawing on recent studies that involved a variety of modelling approaches (*e.g.*, system dynamics or network modelling). In particular, we showed that visual analytics can help modellers during all stages concerning conceptual models, from early design to assessment. As part of this, we highlighted that conceptual models and visual analytics could be mutually beneficial, with a model guiding the search through complex data, leading to the discoveries of new patterns that then strengthen the model. This positive feedback loop presents a clear potential to improve productivity during modelling.

Another avenue that has already been identified as holding potential for increased productivity within modelling and visual analytics focuses on supporting collaborations. Rather than situating collaborations within a specific step of the modelling pipeline (Section 2), our experience suggests that collaboration applies to all parts of the pipeline. Collaboration is indeed a fundamental component of modelling for medical decision-making. In our experience, this exercise requires interdisciplinary teams (*e.g.*, epidemiologists, geneticists, behavioural scientists) and also often involves practitioners, patients, or members of the public with whom findings from models may be shared. While there are techniques such as paired analytics when the analyst and the field specialist can be together, less options are available for large teams that work in several places and on several time zones. Consequently, the integration of modelling and visual analytics should also include support for Distributed and Asynchronous Collaboration (DAC) for interdisciplinary teams. Currently, such support has been centered on offering some of the following features depending on the visual analytics software:

- **Journaling** and **State Tracking**, which are tightly coupled. State tracking is akin to a version control system for all the artifacts in the modeling process. Journaling is a combination of a wiki-like, free-form chronological journal and an automated log file (e.g., including hyperlinks to contextual notes, artifacts) and serves as documentation of the modeling process. Both Journaling and State Tracking establish provenance and reproducibility for various states of the modeling process [7].
- **Annotation** allows the sharing of context for data visualizations and conceptual models. Registering them using journaling can help to structure collaboration.
- **Whiteboarding** is the virtual equivalent of a whiteboard, where documents or visualizations can be shared. This feature exists in many collaborative systems, and it should integrate well with the aforementioned: snapshots of the whiteboard can be managed using state tracking, and annotations should be possible.

To support collaborations in visual analytics *and* modelling requires going beyond these features. By working on a range of medical decision-making projects with interdisciplinary teams, we concluded that an ideal collaboration framework in the medical sciences should be:

- **I**ntuitive. The diverse backgrounds of members in an interdisciplinary team stresses the importance of having a support that is readily usable for all members, without the need of domain specific language or skills. For example, several websites help individuals in managing their personal health records. Anonymised data can be pulled for visualizations, or modelling health behaviours. A collaborative system developed only by modellers may not organize data in ways that make sense to the practitioners in the team, but if the system was only designed for practitioners then medical terms may be present that create a language barrier with modellers. Therefore, setting usability requirements for a wide audience is a challenging but essential task.
- **I**nteractive. A typical way to support asynchronous collaborations is to share visualizations as static products that others can comment on. However, this removes the interactivity that is key to visual analytics. Consequently, a degree of flexibility must be preserved such that visualizations provide a suggested window into the data but do not prevent team members from considering alternatives (e.g., varying parameters of the model and seeing their impact on the visualization).
- **I**ntegrative. Large projects may consist of sub-teams which have produced their own visualizations, using slightly different datasets or modelling assumptions. For example, several regional health authorities may want to look at the 'big picture' regarding the spread of a disease; their models and visualizations could be readily available in reports, but internal regulations prevent a timely sharing of the data. Such situations point out the need to directly integrate visualizations by identifying shared parts, and clarifying competing hypotheses, without accessing the raw data. While this problem has received abundant theoretical treatment in computing science (e.g., as the problem of combining classifiers [12]), the problem of integrating visualizations remains an important obstacle on the way to full collaboration.

These three aspects constitute our 3**I** approach. A software that is 3**I**-compliant would allow users to interact with each other along the various stages of the modelling process in an asynchronous and distributed fashion. We acknowledge that the 3**I** approach requires a leap forward given that only a few software packages currently satisfy fundamental aspects such as journaling or annotating. Nonetheless, we believe that sharing visualizations that are still flexible and supporting the direct integration of visualizations are long-term objectives that are well-worth addressing for the research community.

# 5 Conclusion

Using three case studies, we have demonstrated that visual analytics can be very beneficial to modelling, regardless of the techniques or specific medical problem under study. In particular, we showed that visual analytics can support the conceptual modelling stage and highlighted that its benefits would depend on when it is incorporated. In conclusion, we believe that the coupling of visual analytics and modelling is underway, and that it is time to look further by including the third dimension of collaboration. Our proposed 3**I** approach (intuitive, interactive, integrative) highlights some of the needs and challenges for collaborations, thereby generating essential questions for the development of the next generation of tools.

# 6 Acknowledgments

The authors wish to thank the Vancouver Institute of Visual Analytics and the Modelling of Complex Social Systems program for providing facilities. All authors are indebted to Kyle

Melnick for his visualization of the MEND project, to Grace MacEwan for collecting data on the PHSA report, to Drs Finegood and Matteson for sharing qualitative data regarding MEND, to Andrew Tudwell and Lydia Drasic for their insight about the PHSA report, and to colleagues at Cambridge School of Clinical Medicine for providing feedback.

# References

[1] P. Brantingham, U. Glässer, P. Jackson, and M. Vajihollahi. Modeling criminal activity in urban landscapes. In *Mathematical methods in counterterrorism*, pages 9–31. 2009.

[2] R. Crutzen and P. J. Giabbanelli. Using classifiers to identify binge drinkers based on drinking motives. *Substance use & misuse*, 49(1–2):110–115, 2014.

[3] D. T. Finegood, T. Merth, and H. Rutter. Implications of the foresight obesity system map for solutions to childhood obesity. *Obesity*, 18(S1):S13–S16, 2010.

[4] P. J. Giabbanelli and R. Crutzen. An agent-based social network model of binge drinking among dutch adults. *Journal of Artificial Societies & Social Simulation*, 16(2), 2013.

[5] P. J. Giabbanelli, P. J. Jackson, and D. T. Finegood. Modelling the joint effect of social determinants and peers on obesity among canadian adults. *Theories and Simulations of Complex Social Systems*, pages 145–160, 2014.

[6] P. J. Giabbanelli, T. Torsney-Weir, and V. K. Mago. A fuzzy cognitive map of the psychosocial determinants of obesity. *Applied soft computing*, 12(12):3711–3724, 2012.

[7] D. P. Groth and K. Streefkerk. Provenance and Annotation for Visual Exploration Systems. *IEEE transactions on visualization and computer graphics*, 12(6):1500–1510, 2006.

[8] J. Heer and M. Agrawala. Design considerations for collaborative visual analytics. *Information Visualization*, 7:49–62, 2008.

[9] P. J. Jackson. *A framework for software modelling in social science research*. PhD thesis, Simon Fraser University, Burnaby, Canada, 2013.

[10] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[11] M. Kultys, L. Nicholas, R. Schwarz, N. Goldman, and J. King. Sequence bundles: a novel method for visualising, discovering and exploring sequence motifs. *BMC Proceedings*, 8(S2):S8, 2014.

[12] L. Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley and Sons, 2004.

[13] M. Mokhtari, E. Boivin, and D. Laurendeau. Making sense of large datasets in the context of complex situation understanding. In *Virtual, Augmented and Mixed Reality. Systems and Applications*, volume 8022 of *Lecture Notes in Computer Science*, pages 251–260. Springer, 2013.

[14] S. F. Pratt, P. J. Giabbanelli, and J.-S. Mercier. Detecting unfolding crises with visual analytics and conceptual maps: Emerging phenomena and big data. *Proceedings of the 2013 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 200–205, 2013.

[15] Provincial Health Services Authority. From weight to well-being: time for a shift in paradigms? January 2013.

[16] S. Rossiter, J. Noble, and K. R. Bell. Social simulations: Improving interdisciplinary understanding of scientific positioning and validity. *J. of Artificial Societies & Social Simulation*, 13(1), 2010.

[17] P. Sacher and C. Swain. The MEND Programme: tackling childhood obesity. *British Journal of School Nursing*, 2:4, 2007.

[18] R. G. Sargent. Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation*, pages 130–143. Winter Simulation Conference, 2005.

[19] I. Vandenbroeck, J. Goossens, and M. Clemens. Foresight tackling obesities: Future choices–building the obesity system map. *Government Office for Science, UK Governments Foresight Programme*, 2007.